บทที่ 1

บทนำ

1.1 ความเป็นมาของปัญหา

การวิเคราะห์ข้อมูลทางสถิติบางครั้งมีจุดมุ่งหมายเพื่อทราบความสัมพันธ์ระหว่างตัวแปรสองตัวว่ามีความสัมพันธ์กันในลักษณะใด และการวัดความสัมพันธ์โดยวิธีการทางสถิติมีอยู่หลายวิธี ขึ้นอยู่กับลักษณะของตัวแปรหรือมาตรการวัดค่าตัวแปร

ส่วนความสัมพันธ์จะมากหรือน้อยจะพิจารณาได้จากค่าสัมประสิทธิ์สหสัมพันธ์ประชากร (Coefficient of correlation) สัญญลักษณ์ที่ใช้คือ ρ(rho) ถ้า $|\rho|$ มีค่าใกล้ 1 แสดงว่าตัวแปรคู่ที่กำลังศึกษามีความสัมพันธ์กันมาก ถ้า $|\rho|$ มีค่าใกล้ 0 แสดงว่าตัวแปรคู่ที่กำลังศึกษามีความสัมพันธ์กันน้อย ค่าของ ρ มีเครื่องหมายบวกหรือเครื่องหมายลบ เครื่องหมายของ ρ นี้ จะแสดงทิศทางของลักษณะความสัมพันธ์นั้น

ค่าสัมประสิทธิ์สหสัมพันธ์นอกจากจะทำให้ทราบว่าตัวแปรคู่ใดคู่หนึ่งนั้นมีความสัมพันธ์กันหรือไม่และในทิศทางใดแล้วยังสามารถใช้สำหรับการทำนายหรือคาดคะเนได้อีกด้วย อันจะเป็นประโยชน์ต่อการวางแผนและการตัดสินใจในสภาวะที่ไม่แน่นอน

สัมประสิทธิ์สหสัมพันธ์ระหว่างสองตัวแปร ที่รู้จักกันดีที่สุด คือ สัมประสิทธิ์สหสัมพันธ์เชิงเส้นแบบ Pearson (Pearson Product-Moment Correlation Coefficient) ใช้สัญญลักษณ์ r หรือ r_{XY}

สูตรที่ใช้ในการคำนวน คือ

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left\{N\sum X^2 - (\sum X)^2\right\}\left\{N\sum Y^2 - (\sum Y)^2\right\}}}$$

โดยที่ r คือสัมประสิทธิ์สหสัมพันธ์ตัวอย่างระหว่างตัวแปร X กับตัวแปร Y

$\sum X$ คือผลรวมของค่าข้อมูลของตัวแปร X

$\sum Y$ คือผลรวมของค่าข้อมูลของตัวแปร Y

$\sum XY$ คือผลรวมของผลคูณระหว่างค่าข้อมูลของตัวแปร X และตัวแปร Y

$\sum X^2$ คือผลรวมของกำลังสองของค่าข้อมูลของตัวแปร X

$\sum Y^2$ คือผลรวมของกำลังสองของค่าข้อมูลของตัวแปร Y

n คือจำนวนคู่ของข้อมูล

และ ค่าของ X และ Y จะต้องเป็นค่าที่วัดในลักษณะควบคู่กัน (bivariate (X,Y) data)

<u>ข้อตกลงเบื้องต้น</u> (Assumptions)

1. ตัวแปรทั้งสองต้องเป็นค่าต่อเนื่อง และมีการแจกแจงแบบปกติสองตัวแปร (bivariate normal data]

2. ความสัมพันธ์ระหว่างตัวแปรทั้งสองเป็นแบบเส้นตรง (linear relationship)

3. ข้อมูล (x_i, y_i), $(i=1, .., n)$ เป็นตัวอย่างเชิงสุ่ม

บางครั้งค่าสัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้จากตัวอย่างมีค่าใกล้ศูนย์ ทำให้เกิดข้อสงสัยว่าตัวแปรคู่ที่กำลังศึกษาอยู่นั้นมีความสัมพันธ์กันจริงหรือไม่ ในกรณีเช่นนี้จะต้องทำการทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์ นั่นคือเราอาจทดสอบสมมติฐานใดสมมติฐานหนึ่ง จาก 3 ชนิดคือ

1. $H_0 : \rho = 0$ $\quad H_1 : \rho \neq 0$
2. $H_0 : \rho \geq 0$ $\quad H_1 : \rho < 0$
3. $H_0 : \rho \leq 0$ $\quad H_1 : \rho > 0$

ตัวสถิติที่ใช้ในการทดสอบนี้คือ

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

แต่กรณีที่ต้องการทดสอบสมมติฐาน

1. $H_0 : \rho = \rho_0$ $\quad H_1 : \rho \neq \rho_0$ $\quad (\rho_0 \neq 0)$
2. $H_0 : \rho \geq \rho_0$ $\quad H_1 : \rho < \rho_0$ $\quad (\rho_0 \neq 0)$
3. $H_0 : \rho \leq \rho_0$ $\quad H_1 : \rho > \rho_0$ $\quad (\rho_0 \neq 0)$

ตัวสถิติที่ใช้ในการทดสอบข้างต้นใช้ไม่ได้ จำเป็นต้องใช้ Fisher's transformation

R.A Fisher เป็นผู้ค้นพบการกระจายค่า r ในปี ค.ศ. 1915 ซึ่งการแจกแจงของ r ขึ้นอยู่กับค่าของ ρ และ n เท่านั้น การแจกแจงของ r จะสมมาตรเมื่อ $\rho = 0$ แต่จะมีลักษณะเบ้ถ้า $\rho \neq 0$ โดยจะเบ้ซ้ายถ้า $\rho > 0$ และเบ้ขวาถ้า $\rho < 0$ ซึ่งถ้า ρ มีค่าเป็นบวกค่าของ r อาจจะให้ค่าที่ต่ำกว่า ρ แต่ถ้า ρ มีเครื่องหมายลบค่าของ r อาจจะสูงกว่าค่าของ ρ ได้

เนื่องจากความเบ้นี้เอง Fisher จึงได้ดัดแปลงข้อมูลเสียใหม่ โดยใช้ ln (logฐาน e หรือ natural log) ดัดแปลงข้อมูลเดิมจาก r มาเป็น z_F โดยที่

$$z_F = \tfrac{1}{2} \ln \frac{1+r}{1-r} = 1.1513 \log \frac{1+r}{1-r}$$

การแจกแจงของ z_F นี้จะประมาณได้ด้วยการแจกแจงปกติ มีค่าเฉลี่ย

$$\mu_{z_F} = \tfrac{1}{2} \ln \frac{1+\rho}{1-\rho} = 1.1513 \log \frac{1+\rho}{1-\rho}$$

และมีส่วนเบี่ยงเบนมาตรฐาน

$$\sigma_{z_F} = \frac{1}{\sqrt{n-3}}$$

ดังนั้นตัวสถิติที่ใช้ในการทดสอบคือ

$$z = \frac{\tfrac{1}{2} \ln \left(\frac{1+r}{1-r}\right) - \tfrac{1}{2} \ln \left(\frac{1+\rho}{1-\rho}\right)}{1/\sqrt{n-3}}$$

เนื่องจากความถูกต้องของการประมาณค่าขึ้นอยู่กับขนาดของตัวอย่าง ปัญหาก็คือขนาดตัวอย่างเท่าไรจึงจะเหมาะสมที่ทำให้การดัดแปลงข้อมูลโดยวิธีนี้ได้ผลดี นั่นคือทำให้ค่าความผิดพลาดที่เกิดจากการปฏิเสธสมมติฐานเมื่อสมมติฐานนั้นถูกต้อง ที่เกิดขึ้นจริงจากการทดลองใกล้เคียงหรือเท่ากับค่า α ที่กำหนดให้ในการทดสอบค่า ρ ที่ระดับต่าง ๆ

1.2 วัตถุประสงค์ของการวิจัย

วิทยานิพนธ์เรื่องนี้มีวัตถุประสงค์ที่จะ

1. สร้างตารางข้อมูลชนิดไบวารีเอทนอร์มอล โดยวิธีการซีมูเลชั่น
2. ตรวจสอบลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์ตัวอย่างเมื่อ $\rho \neq 0$
3. ตรวจสอบลักษณะการแจกแจงของค่า Fisher's transformation
4. เพื่อหาขนาดตัวอย่างที่เหมาะสมที่ทำให้ การดัดแปลงข้อมูลโดยวิธี Fisher's transformation ได้ผลดี ในการทดสอบสมมติฐาน 3 แบบคือ

   1. $H_0 : \rho = \rho_0 \qquad H_1 : \rho \neq \rho_0 \qquad (\rho_0 \neq 0)$
   2. $H_0 : \rho \geq \rho_0 \qquad H_1 : \rho < \rho_0 \qquad (\rho_0 \neq 0)$
   3. $H_0 : \rho \leq \rho_0 \qquad H_1 : \rho > \rho_0 \qquad (\rho_0 \neq 0)$

1.3 ขอบเขตของการวิจัย

เนื่องจากการศึกษาเรื่องสัมประสิทธิ์สหสัมพันธ์เราได้ตั้งข้อสมมติไว้ว่าความสัมพันธ์โดยเฉลี่ยระหว่างตัวแปรเหล่านั้นมีลักษณะเป็นเส้นตรง ที่เกี่ยวข้องกับการแจกแจงของค่าตัวแปรสองตัว (Bivariate distribution) ซึ่งได้มาโดยการสุ่มจากประชากร

ดังนั้นในการศึกษาเรื่องนี้จะศึกษาเฉพาะข้อมูลที่มีการแจกแจงแบบปกติสองตัวแปร (The bivariate normal distribution)

1.4 ประโยชน์ที่จะคาดว่าจะได้รับ

1. สามารถนำตารางตัวแปรสุ่มที่มีการแจกแจงแบบปกติสองตัวแปร ไปใช้ประโยชน์ในการศึกษาเรื่องอื่น ๆ ที่ต้องใช้ข้อมูลประเภทนี้ได้

2. เพื่อใช้เป็นแนวทางสำหรับนักวิจัย ที่ต้องการหาขนาดตัวอย่างที่เหมาะสมสำหรับการวิเคราะห์ค่า สัมประสิทธิ์สหสัมพันธ์ได้อย่างมีประสิทธิภาพ ณ ระดับนัยสำคัญ 3 ระดับ คือ 0.1, 0.05 และ 0.01

3. เพื่อเป็นแนวทางสำหรับการหาวิธีการแปลงค่าข้อมูล เพื่อให้ได้ค่าสถิติเพื่อการทดสอบสัมประสิทธิ์สหสัมพันธ์ที่มีคุณสมบัติตัวสถิติที่เหมาะสม

1.5 คำสำคัญ (Key word) สำหรับเรื่องนี้คือ

สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient)

ซีมูเลชั่น (Simulation)

การแจกแจงแบบปกติสองตัวแปร (The Bivariate normal distribution)

เลขสุ่ม (Random number)

การแจกแจงสม่ำเสมอ (Uniform distribution)